Demonstrate methods to generate random numbers for the probability distributions.

Use appropriate data wrangling techniques to create a dummy data frame containing two columns with python.

OR

Explain the methods to divide a data set into training and test data sets.

Employ data wrangling techniques to append two data sets with python.

Discuss the important properties of ROC curve and outline the steps to plot an ROC curve.

Differentiate the linear regression model with scikit-learn and statsmodel library.

OR

**Describe the following:**

**i) Variance Inflation Factor ii) Residual Standard Error iii) Linear Regression**

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis.

Multicollinearity inflates the variance and type II error. It makes the coefficient of a variable consistent but unreliable.

It estimates for the inflation in the regression coefficient caused due to multicollinearity.

VIF measures the number of inflated variances caused by multicollinearity.

The VIF needs to be calculated for each of the variables and if the value is very high for a particular variable, then that predictor needs to be eliminated from the model.

The following process goes under the hood for calculation of VIF:

Write Xi as a linear function of other predictor variables:

https://lh4.googleusercontent.com/qt_FzblYxoWr9KMQPorRCW62GRAQ9qh5fIlFvp_zlSQuwgGsJtrNuBRW_M308wrADyWLXiM4S3WMsE63VqqlV1wBEEM849PbwLzz7bEcvZpXY0Irekmn0XBMJy_h5inrWyZ0brRv_DP3E0m5W5yuD0rnn494L_1I1eS-CKPHVABgQB0NUjwnMjgCbeXqJA

Calculate the coefficient of determination for this model and call it R2. The VIF for Xi is given by:

https://lh5.googleusercontent.com/41O15NV-YjY0ScS0a3Q2ZZjsMUuzLJDBnbE-SbJokhttW_ghuZR9-BoRvbBn3s6u0DYDoVJ66pCbjlL_OuZBGMFAvrPExYo3LXIfaPtzCxVngOM7QK_Pb8fhXdb2nWvlT-bUCt3MObrCh40mXEQT1xkshE0M_n_2H3Q-RP9nd4JmlWf763ap0MM8OJaMjQ

If the VIF=1, then the variables are not correlated. If 1<VIF<5, then the variables are moderately correlated with other predictor variables and can still be part of the model. If VIF>5, then variables are highly correlated and need to be eliminated from the model.

**ii) Residual Standard Error:**

Another concept to learn is the concept of Residual Standard Error (RSE). It is defined as:

https://lh5.googleusercontent.com/NuK6f-bnRo5XwaQs7xyUnhU9fRihx-eY56yZGO9BHzabIPrBgNDMFgOQMu3EOsAgJNGuNBHBT6AXtyD3CEeBpDQMxJXvm2xpnUvtBVADM2-tbPr9_gl4zcG6M82fpCbquWg_0eX4x8Qvcz0jHDuU8PX18uUHlWNHUhMPmFxzJNNrkN3nI6uJEwsP-E4cgwand So, RSE can be written as

https://lh5.googleusercontent.com/iX-x9pgBtFQmwxXwK2uNqtNPQKa9IQ8NyvZ_CwfNwBH5ozrJWhJoGTKm_Xn0GJLz2Q4pFLCUEloZswOPeCOTX_3BwFAKBWUGWlBV7E2pGR44nB_PR8PvcrpWp7Eru97qMgMFtEKMqNDgj3Zlm_3ubaBHH6G_qRtrTdT5TT_goLkl4MpTbv2rcguMeeWvfQa simple linear regression model.

Where n=number of data points.

In general,  https://lh6.googleusercontent.com/MjVaWWhXJynDjtEv0XjZNl7eSt_97szUI43C8xxIM4OdsMv5f72CLIozzOg2hHKOvyFuOQu7OHJQuddpBsf1oFd_9yzcKYUJt4xrc2h-vkRkbtOYpqP-JsIeO4wLsCiaswuTs2qjg0Ebg1U5hvhDwgPf8pIei830ljNCppDJOw-KFSh7LtuU2Uq6ydri3gwhere p=number of predictor variables in the model.

The RSE is an estimate of the standard deviation of the error term (res). This is the error that is inevitable even if the model coefficients are known correctly.

In multiple regressions, values of the RSE generally go down, as we add more variables that are more significant predictors of the output variable).

**iii) Linear Regression:**

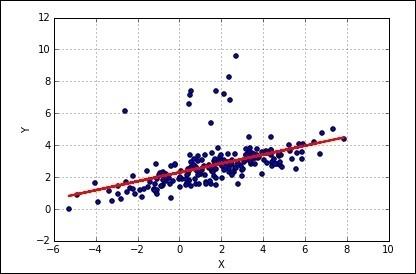
Linear regression quantifies the relationship between one or more predictor variable(s) and one outcome variable. Linear regression is commonly used for predictive analysis and modeling. For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable). Linear regression is also known as multiple regression, multivariate regression, ordinary least squares (OLS), and regression.

**Demonstrate visualization techniques to represent outliers in the data set and use the best fit line developed from the model with outliers.**

Outlier is an observation that appears far away and diverges from an overall pattern in a sample.

In statistics, an outlier is an observation point that is distant from other observations.

The outliers need to be removed or properly treated before using the dataset for modeling. The outliers can distort the model and reduce its efficiency even if they are less in number, compared to the size of the dataset. As low as 1% outlier data is also capable enough to distort the model. It is actually not the number of outlier points but the degree to which it is different from an average point that determines the degree of distortion.



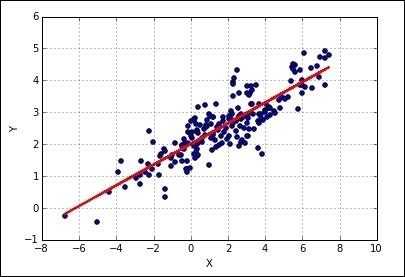
Model with outliers

**Detect Outliers:**

* Most commonly used method to detect outliers is **visualization**.
* We use various visualization methods, like **Box-plot, Histogram, Scatter Plot.**
* Some analysts also use various **thumb rules** to detect outliers.
  + Any value, which is **beyond the range of Q1-1.5 x IQR to Q3+1.5 x IQR**.
  + Use capping methods. Any value which is out of **range of 5th and 95th percentile** can be considered as outlier.
  + Data points, **three or more standard deviations away from mean** are considered outliers.

**Remove Outliers:**

* **Deleting observations:** We delete outlier values if it is due to data entry error, data processing error or outlier observations are very small in numbers.
* **Transforming and binning values:** Transforming variables can also eliminate outliers. Natural log of a value reduces the variation caused by extreme values. Binning is also a form of variable transformation. Decision Tree algorithm allows to deal with outliers well due to binning of variables. We can also use the process of assigning weights to different observations.
* **Imputing:** Like imputation of missing values, we can also impute outliers. We can use mean, median, mode imputation methods.
* **Treat separately:** If there are a significant number of outliers, we should treat them separately in the statistical model. One of the approaches is to treat both groups as two different groups and build individual models for both groups and then combine the output.



Model without outliers

Illustrate Decision Tree classifier that can classify a flower as belonging to a certain species based on the flower petal and sepal dimensions with python.

Paraphrase cross-validating and pruning the decision tree.

OR

Describe the following: i) Contingency table ii) Conditional probability.

Demonstrate logistic regression with python for the following Boston house price sample dataset.

